CASTIGO ALTRUISTA: REVISIÓN DEL CONCEPTO.

Carlos Ramos Aguirre

Laboratorio de Sistemática Humana,
Universidad de las Islas Baleares.

Castigo altruista (*altruistic punishment*) es la sanción que un individuo cooperador aplica a otro no-cooperador o tramposo (*free-rider*) pese a que tal acción punitiva conlleva también un coste para si mismo. Esta peculiar modalidad de sanción ha sido interpretada como una conducta de tipo altruista, garante de la cooperación en los grupos humanos. En el presente trabajo se ofrece una revisión crítica del concepto desde el momento de su aparición hasta su estado actual.

Palabras clave: castigo altruista; altruismo; castigo costoso; cooperación humana.

1. INTRODUCCIÓN.

Seguramente la forma más adecuada de explicar en qué consiste el castigo altruista sea comenzando por decir que se trata de una forma de altruismo o conducta altruista. Pero, como ocurre siempre que un concepto es compartido por varias disciplinas, la falta de delimitación lo bastante nítida de los diferentes significados, puede contribuir a crear cierto estado de confusión. Este es el caso del altruismo (o conducta altruista) que, bien merece una aclaración previa. El término *altruismo* posee distintos significados según se le aborde desde la biología, la psicología, la filosofía o la sociología y, no obstante, al menos estas cuatro disciplinas lo han incorporado a sus respectivos discursos. Cada una con su sentido y su matiz, que trataré de mostrar de inmediato.

Desde la perspectiva biológica el altruismo sólo toma sentido en un contexto de selección natural ya sea individual o de grupos. La idea base que barajan muchos científicos desde la década de 1960 es que, de algún modo, los efectos positivos del altruismo sobre la selección de grupos compensan los efectos negativos que éste tiene sobre la selección individual. La explicación de cómo se da esa compensación todavía es algo que se discute hoy día. Disciplinas como la Biología, la Etología, o la Genética de poblaciones definen el

altruismo en términos exclusivamente de reproducción y supervivencia. Así el *altruismo biológico* es todo acto que aumenta la aptitud (*fitness*) de los demás a costa de la aptitud del propio actor (Sober y Wilson; 2000). En el tantas veces citado *Sociobiología: la nueva síntesis* de Edward O. Wilson aparece una definición de altruismo que marcha en la misma línea de la propuesta anterior, pero acaso sea más expresiva: altruismo es el comportamiento auto-destructivo ejecutado en beneficio de otros (Wilson; 1980). En opinión de algunos autores este giro dramático es absolutamente necesario para hablar de un auténtico altruismo biológico.

Para la psicología, en cambio, es otro el significado que se le atribuye al denominado *altruismo psicológico*, ya que el enfoque se dirige hacia el campo de las motivaciones y no al de las conductas. De esta forma el altruismo se manifestaría en algunas -no en todas- de las acciones que determinados individuos perpetran teniendo como objetivo remoto el bienestar ajeno. Es decir, que un individuo ayuda a los otros por ellos mismos y no por el beneficio que eventualmente pueda obtener, y que tal conducta tiene un origen psicológicamente remoto. El hecho de que una acción se identifique como altruista psicológica no obsta para que pueda también ser considerada altruista biológica, pero ambas consideraciones, que pueden ser coincidentes, responden a dos realidades distintas (Sober y Wilson; 2000).

Por su parte el *altruismo moral* o *altruismo humano* es un concepto estudiado tradicionalmente por la Filosofía Moral y la Filosofía Política. En buena medida su definición se solapa con la de altruismo psicológico en tanto que atiende a las motivaciones remotas. Algunos autores consideran que el altruismo moral es, de algún modo, un tipo de altruismo biológico ya que las conexiones entre ambos son innegables. Sin embargo, hay que aclararlo, esto no significa que al hablar de altruismo en los insectos sociales, por ejemplo, y de altruismo en los seres humanos, nos estemos refiriendo al mismo fenómeno (Cela Conde, 1985; Cela Conde, ¿?; Cela Conde, 2005).

En el ámbito social, el estudio del altruismo -entendido éste como filantropía-, se encuentra entroncado en un primer momento con el propio nacimiento de la ciencia sociológica, hacia mediados del siglo XIX y coincidiendo con la aparición de la era industrial. Además fue Auguste Comte (1798-1857), creador de la palabra híbrida *sociología*, quien también introdujo el término *altruismo* <<para referirse a un tipo particular de comportamiento moral por el cual una persona intenta hacer el bien a los otros sin pensar en ninguna ventaja que pudiera derivarse>> (Turbón, 2006). Más modernamente la concepción filantrópica ha dado paso a una concepción con fuerte carga cultural, siendo el *altruismo social*

relacionado con fenómenos como la cooperación y la cohesión del grupo. El pensador francés Emile Durkheim (1858-1917), considerado uno de los padres de la sociología, también habló de *suicidio altruista* como uno de los tipos de suicidio, que se daría por ejemplo cuando con la muerte del suicida se evitan maldiciones al resto de la comunidad o, sencillamente, cuando la muerte del suicida es lo socialmente correcto ya sea por tradición, religión o derecho.

El estado de relativa confusión al que aludo al comienzo de este epígrafe viene propiciado por la falta de conciencia acerca de los diferentes significados que, en síntesis, acabo de presentar. En términos lingüísticos podríamos decir que tenemos un único significante para al menos cuatro significados diferentes. En definitiva, nos encontramos ante el conocido problema de la polisemia. Pero conviene tener esta circunstancia presente a la hora de aproximarnos a la literatura científica sobre el altruismo y el castigo altruista, para no realizar saltos inverosímiles entre unas perspectivas y otras.

2. HISTORIA DEL CONCEPTO.

2. 1. ANTECEDENTES.

La aparición de la teoría evolutiva en el panorama científico mundial o, de modo más exacto, la enunciación suficientemente madura de ésta, cambió de forma definitiva -siguiendo la conocida sentencia de Theodosius Dobzhansky-[1] el modo en que debe pensarse la biología. Charles Darwin (1809-1882) es, con justicia, la figura más celebrada de esta corriente que supuso una auténtica revolución científica, pero sin duda es larga la lista de predecesores y sucesores que también merecen ser nombrados. No lo haremos aquí, pues escapa a los límites de nuestra tarea, aunque el nombre de algunos de ellos se irá deslizando por necesidad en el transcurso de este trabajo. Darwin muere en 1882 dejando tras de si un corpus teórico muy completo materializado en una selecta obra, que sigue siendo admirada y admirable casi ciento cincuenta años después de su aparición, y que es fruto de sus experiencias y reflexiones acerca de las diversas formas de vida. *El origen de las especies* (1859) y *El origen del hombre y la selección en relación al sexo* (1871) son, en opinión generalizada, sus escritos cumbre, aunque cabría destacar otros textos entre los que se encuentra *La expresión de las emociones en el hombre y en los animales* (1873).

[1] <<En la biología nada tiene sentido si no se considera bajo el prisma de la evolución>> citado en Ayala (1999).

Por otro lado, el balance de la influencia que el evolucionismo ejerció sobre las ciencias sociales fue, en tiempos de Darwin, más negativo que positivo. En verdad ciencias naturales y ciencias sociales nunca han sido compartimentos estancos, tampoco en el siglo XIX. Es bien sabido que Darwin recibió, a la hora de concebir el concepto fundamental de selección natural, influencia de los escritos sobre población de Thomas R. Malthus (1766-1834). También debemos recordar que, al tiempo que Darwin escribía sus obras, Herbert Spencer (1820-1903) hilvanaba sus teorías sociales en la línea de lo que dio en llamarse *Organicismo*[2]. Aunque al parecer Darwin no estimaba demasiado a Spencer, en lo que respecta a intelectualidad, tomó de éste la expresión <<la supervivencia del más apto>>, que incluyó en las sucesivas ediciones de *El origen de las especies*. El tipo de ideas que propugnó Spencer se han conocido en la historia con el nombre de *darwinismo social* y a menudo esta corriente es considerada por los sociólogos y antropólogos (culturales) como una mácula en el historial de las ciencias sociales. Eso sí, una mácula de la que han escarmentado. El fenómeno de transplantar conceptos de la biología a las ciencias sociales tuvo algunas repercusiones nefastas. Durante las últimas décadas del siglo XIX y primeras del XX, los argumentos del darwinismo social sirvieron para legitimar discursos y políticas de corte racista. Además, en el campo de la criminología se cometieron crueles errores al amparo de ideas como la del *criminal atávico* de Cesare Lombroso (1835-1909). Esta clase de hechos han conducido a una considerable degradación de los canales que tradicionalmente han servido de puente entre las ciencias naturales y sociales. Tal vez, en la actualidad, con los crecientes avances de las ciencias cognitivas y las neurociencias pueda pronto reestablecerse una comunicación que parece imprescindible, por el enriquecimiento mutuo que suele suponer.

Por lo que hace a este trabajo, trataré de presentar ahora las relaciones existentes entre el evolucionismo y el *problema* del altruismo. En realidad, el problema del altruismo se presenta tempranamente. Según sostiene la teoría evolutiva, los organismos de todas las especies actúan, en sus respectivos hábitats, maximizando su aptitud individual, para así garantizar la continuidad de la especie. Pero puesto que la cantidad de recursos de un hábitat es limitada, como consecuencia, resultan excedentes de población. Esto es, individuos que no podrán acceder a los recursos necesarios para la supervivencia. Este hecho se traduce en la necesidad de competir por el alimento y (en el caso de los machos) por las

[2] Organicismo es la comparación de la estructura y funcionamiento de la sociedad con la estructura y funcionamiento de los organismos biológicos; lleva consigo un paralelismo detallado entre los sistemas de nutrición, comunicación, transporte, etc., y con los sistemas de estructura-función de los animales (Pratt Fairchild, 2001).

hembras. En esta competición, que no se realiza con garras y dientes necesariamente, algunos organismos perecen y otros sobreviven, los más aptos, que transmitirán sus cualidades a la siguiente generación. El mecanismo externo que actúa seleccionando a unos organismos y no a otros es la selección natural. Hasta aquí, todo correcto. Pero Darwin pronto se da cuenta que en algunas especies, hay individuos que se comportan de un modo anómalo o antinatural en tanto que, en determinadas situaciones, reducen su aptitud individual en favor de la aptitud de otros.

El asunto del altruismo lo trata Darwin en *El origen del hombre y la selección en relación al sexo* (1871) y, concretamente, entre los capítulos III y V. Darwin intentó explicar la conducta moral humana desde un enfoque evolucionista. Sabía de la diversidad de códigos morales en los distintos pueblos pero creía en la existencia de una base común. Dicha base la encuentra en el denominado sentido moral (*moral sense*) que, a un tiempo, sería común a todos los humanos y extraño a todos los no-humanos. A partir del sentido moral se habrían ido construyendo las distintas sociedades, en el seno de las cuales irían surgiendo distintos códigos éticos que ocuparían los distintos peldaños de una escalera hacia el progreso moral. Este tipo de concepción es heredada de la idea ilustrada de la perfectibilidad del hombre. Pero Darwin, además, anticipó la idea de selección de grupos: <<al ser incapaz de dar una explicación al comportamiento ultrsocial de los himenópteros, habló de las ventajas adaptativas que tendría un grupo de cooperadores frente a otro de individuos egoístas>> (Ayala y Cela Conde, 2006).

La cooperación humana es con seguridad la expresión más notable de altruismo observada entre seres vivos, aunque no por ello carente de discusión. Pero la ciencia sólo recientemente ha considerado el altruismo o la conducta cooperativa como objetos de estudio a tener en cuenta. Salvando algunos casos aislados, no será hasta la década de 1960 que los científicos dirijan un considerable caudal de sus esfuerzos a investigar el motivo por el que los individuos de algunas especies son capaces de sacrificar incluso sus vidas en pro de otros, habitualmente congéneres con quienes guardan parentesco, aunque no necesariamente sea siempre así. El altruismo se presenta entonces como el centro de la escena científica para disciplinas hasta entonces no muy desarrolladas como la etología, la genética de poblaciones u otras ciencias de la conducta.

En 1962 Wynne-Edwards publica *Animal dispersion in relation to social behaviour*, en donde da una explicación a la conducta altruista de algunos animales a través de la *selección de grupos*. Wynne-Edwards es el primero en plantear de un modo formal el mecanismo de selección grupal. Observa que en algunas especies, cuando se alcanza un volumen de población que pone en peligro la

supervivencia del grupo por existir riesgo de sobreexplotación del hábitat, determinados individuos "altruistas" renuncian a su capacidad reproductiva. Este mecanismo, la selección de grupos, beneficia a la especie pero no necesariamente al individuo y de ahí su nombre. Wynne-Edwards llega a plantear la posibilidad de que todo comportamiento social tenga origen en las exhibiciones *epideícticas* que realizan los individuos de una especie, cuya función sería informar al grupo sobre la densidad de la población (Smith, 1984; Cela Conde y Ayala, 2005). Pero se plantea un problema importante; si los individuos altruistas de un grupo no se reproducen, sus genes no pasan a la siguiente generación. No puede ser por tanto una estrategia evolutivamente estable.

El modelo de Wynne-Edwards acusa algunas taras y es rechazado, quedando el concepto de selección de grupos en hibernación hasta que años más tarde sea recuperado por Elliott Sober y David Sloan Wilson. Pero en 1964 W. D. Hamilton propone su modelo de *selección de parentesco*. La explicación que Hamilton encuentra al fenómeno del altruismo consiste en que los actos altruistas de un individuo tienen sentido cuando están orientados a incrementar la aptitud de un pariente. Cuanto más próximo es el parentesco entre dos sujetos, mayor es la cantidad de genes que comparten y, por tanto, mayor probabilidad de ser ayudado por el otro sujeto. La idea germinal es que los genes "buscan" la mejor forma de perpetuarse de una generación a otra. A partir de Hamilton el asunto del altruismo y el egoísmo toma relevancia y en la década de los 70 son varios los científicos que logran gran popularidad gracias a este enfoque. Cabe mencionar, entre otros, a John Maynard Smith, Edward O. Wilson[3] y Richard Dawkins[4]. No obstante la selección de parentesco también presentaba algunas dudas. Por ejemplo, se objetó que para poder cooperar con un pariente, primero es preciso tener la capacidad para poder reconocerlo como tal. Un escollo, por tanto, nada desdeñable para muchas especies. Pero, además, la selección de parentesco dejaba sin explicar la cooperación entre individuos genéticamente no relacionados, que era observable en varias especies.

Así, con estas incógnitas pero con un creciente interés por los fenómenos de la cooperación y el altruismo, en 1971 Robert L. Trivers publica un artículo, clásico ya en todas las bibliografías sobre la materia, en que presenta el concepto de *altruismo recíproco* (Trivers, 1971). El altruismo recíproco es un mecanismo basado, como su propio nombre indica, en la reciprocidad. Es el <<yo te ayudo a ti y tú me ayudas a mi>>. Es un mecanismo muy válido para

[3] Con sus polémicos libros *Sociobiología. La nueva síntesis* (1975) y *Sobre la naturaleza humana* (1978).
[4] Con su muy difundido libro *El gen egoísta* (1976).

explicar, por ejemplo, la cooperación en los primates. Aunque Trivers lo concibió como un mecanismo cooperativo incluso interespecies. El altruismo recíproco es a veces referido como reciprocidad directa para distinguirlo de la reciprocidad indirecta.

Pero el altruismo recíproco exige la existencia de redes de relación duraderas en el tiempo, donde se dan interacciones repetidas, y donde el sujeto que en un momento dado presta ayuda a otro, esperará que éste se la devuelva, si así lo necesitarse. Aunque el poder explicativo del altruismo recíproco es grande, en sociedades complejas y muy numerosas, como las humanas, los individuos se ven obligados diariamente a interactuar y a cooperar con muchos otros con los que es probable que no se repita interacción. Pero con todo sigue dándose cooperación, y esto merecía continuar siendo investigado. En 1987 R. D. Alexander propone el mecanismo de la *reciprocidad indirecta* que se basa en la posibilidad de formar reputaciones. Este mecanismo exige de entrada, como podrá advertir el lector fácilmente, organismos con un desarrollo cognitivo bastante avanzado. Es decir, los humanos y tal vez -aunque es poco probable- algún otro primate con un alto grado de encefalización. La reciprocidad indirecta se da cuando la conducta cooperadora de un individuo hacia otro puede propiciar que, a su vez, un tercero se sienta confiado a cooperar con él. O lo que es lo mismo, que la conducta benévola de una persona aumenta las posibilidades de recibir ayuda de otras.

Pese a todo, los tres mecanismos que acabamos de revisar, a saber: la selección de parentesco (o selección familiar), el altruismo recíproco (o reciprocidad directa) y la reciprocidad indirecta (o mecanismos basados en la reputación); no satisfacen completamente la explicación de la cooperación. Ya que, en los humanos se puede observar conducta altruista entre individuos no relacionados genéticamente, que interactúan una sola vez y en condiciones en que no es posible forjar reputación. Pues bien, la propuesta que viene a completar la gama de explicaciones ya existentes al fenómeno de la cooperación, y en concreto de la cooperación humana, aparece en 2002 de la mano de dos investigadores suizos, Ernst Fehr y Simon Gächter, y se denomina *castigo altruista*.

Cuadro resumen de las teorías explicativas del altruismo.

TIPO DE ALTRUISMO	RELACIÓN GÉNICA ENTRE INDIVIDUOS	Nº MÍNIMO DE INTERACCIONES	ÁMBITO EVOLUTIVO	AUTOR
PARENTESCO / GENES	GENÉTICAMENTE RELACIONADOS	INTERACCIÓN CONTINUA	SELECCIÓN DE PARENTESCO	HAMILTON (1964)
RECIPROCIDAD DIRECTA	NO RELACIONADOS	MÁS DE UNA (con el mismo sujeto)	ALTRUISMO RECÍPROCO	TRIVERS (1971)
RECIPROCIDAD INDIRECTA	NO RELACIONADOS	MÁS DE UNA (con distintos sujetos)	SELECCIÓN DE GRUPO	ALEXANDER (1987)
CASTIGO ALTRUISTA	NO RELACIONADOS	UNA SÓLA	SELECCIÓN DE GRUPO	FHER Y GÄCHTER (2002)

2. 2. APARICIÓN Y DESARROLLO.

El concepto de castigo altruista (*altruistic punishment*) resulta de reciente acuñación. Se enmarca dentro de las teorías construidas al objeto de explicar la conducta cooperativa de las diversas especies y en particular de la especie humana.

En septiembre del año 2000 un catedrático de Microeconomía y Economía Experimental de la Universidad de Zurich, Ernst Fehr, y su colega Simon Gächter publican un artículo en *The American Economic Review* en el que explican los resultados que han obtenido en un experimento sobre bienes públicos (Fehr y Gächter; 2000). Dichos autores demuestran de forma experimental que, ciertamente, los individuos cooperadores presentan una marcada tendencia a castigar a los tramposos. Y añaden que tal tendencia se cumple incluso aunque el castigo sea costoso y no conlleve ningún beneficio material para el castigador. Además sostienen que, a tenor de lo revelado por sus experimentos, los tramposos son castigados más severamente cuanto más se alejan del nivel de cooperación de los cooperadores. Así las cosas, un individuo tramposo podría evitar el castigo o al menos reducirlo incrementando su cooperación. Fehr y Gächter terminan sugiriendo que, en aquellos casos en que existe oportunidad para el castigo, el nivel de trampas es menor (Fehr y Gächter; 2000).

Creo que el artículo arriba referido constituye la primera presentación de lo que en esencia es el castigo altruista (*altruistic punishment*), salvo por un detalle nada despreciable, que en ningún momento del texto aparecen unidas las palabras "castigo" y "altruista". O lo que es lo mismo, que se ha dado con la cosa, pero todavía no se le ha nombrado.

Entre tanto, en un artículo escrito con anterioridad a la aparición formal del concepto de castigo altruista y publicado en 2001 en el *Journal of Theoretical Biology*, Joseph Henrich y Robert Boyd proponen un modelo verosímil capaz de articular la cooperación humana y el castigo a los tramposos (Henrich y Boyd., 2001). El primer obstáculo con que se topan ambos investigadores es el dilema de segundo orden que generan aquellos individuos cooperadores que sin embargo no están dispuestos a castigar a los tramposos. De este modo, los cooperadores que sí castigan están proveyendo al grupo de un bien público y los cooperadores que no castigan pasan a ser tramposos de segundo orden (*second-order free-riders*). Esta circunstancia, como el lector advertirá, puede repetirse sucesivamente hasta un número indefinido de órdenes. Henrich y Boyd resuelven este dilema razonando que mediante dos mecanismos de tipo psicológico como son 1) copiar la conducta de la mayoría y 2) copiar la conducta más exitosa, los individuos pronto asumen una actitud cooperativa y castigadora. Esto limita el dilema a un número finito de órdenes. Y continúan los autores afirmando que, una vez estabilizada la cooperación, mediante una forma particular de selección cultural de grupo, este rasgo beneficioso se transmite a otras poblaciones humanas.

La explicación de la cooperación humana mediante el mecanismo del castigo llevaba masculándose algunos años. Este hecho no tiene nada de particular ya que desde antiguo, en las ciencias sociales, se venía hablado de las fuerzas e instituciones coercitivas como las responsables de soportar el orden social. No es extraño, pues, que esta explicación terminase migrando a las ciencias biológicas.

En 2002, aunque las investigaciones ya venían de atrás, Michael E. Price, Leda Cosmides y John Tooby, investigadores todos ellos de la Universidad de California, publican un artículo que vendría a proponer el sentimiento punitivo (*punitive sentiment*) como un mecanismo psicológico anti-tramposos (Price; Cosmides; Tooby; 2002). Se trata de un interesante artículo con elevado valor didáctico que ronda, sin formularlo, el concepto de castigo altruista. Los autores sugieren que los sentimientos punitivos que en ocasiones sienten los participantes de un trabajo colectivo hacia otros -menos participantes-, podrían ser adaptaciones para solventar problemas relacionados con la denegación de ayuda en tareas colectivas. Si esto

es así, el sentimiento punitivo cumpliría dos funciones principales, a saber: a) reclutamiento de trabajo para las acciones colectivas; b) eliminación de las diferencias o hipotéticas ventajas adaptativas que proporcionaría la adopción de la estrategia del tramposo.

Pero en enero de 2002 Fehr y Gächter vuelven a publicar un artículo juntos, esta vez en la revista *Nature*, y que lleva por título una por entonces inédita expresión: *Altruistic punishment in humans*. Ésta es, considero, la primera ocasión en que el castigo altruista es denominado como tal (Fehr y Gächter; 2002). En este artículo se dice -o más bien se repite- que <<castigo altruista significa que los individuos castigan, aunque el castigo sea costoso para ellos y no les proporcione beneficio material alguno>> y que la cooperación florece allí donde el castigo altruista es posible y se debilita allí donde no lo es. Además vaticinan los autores que cualquier estudio futuro acerca de la evolución de la cooperación humana deberá centrar sus esfuerzos en explicar el castigo altruista. Mas hay algo nuevo en este artículo de 2002 que no se encuentra en el de 2000. Los autores se han cuestionado qué es lo que motiva ese acto aparentemente irracional que es el castigo altruista y responden levemente diciendo que el castigo altruista debe ser promovido por las fuertes emociones negativas (rabia) que los tramposos provocan en los cooperadores.

El concepto de castigo altruista ha sido acuñado públicamente en 2002, es cierto, pero por entonces todavía están lejos de cumplirse los vaticinios de Fehr y Gächter. En realidad existe un marasmo de investigadores tras el asunto de la cooperación entre seres humanos, que tratan de apuntalar lo que se escapa a los planteamientos teóricos de la selección de parentesco (Hamilton, 1964), el gen egoísta (Dawkins, e.o. 1976; 2005), el altruismo recíproco (Trivers, 1971) o la reciprocidad indirecta (Alexander, 1987)[5]. Robert Boyd, Joseph Henrich, Peter J. Richerson, Michael E. Price, Leda Cosmides y John Tooby, como se ha visto, son algunos de los científicos que centran parte de sus esfuerzos en estudiar las conexiones entre cooperación y castigo (Boyd y Richerson., 1992; Richerson y Boyd, 1997; Henrich y Boyd., 2001; Price, Cosmides y Tooby; 2002). Gran número de las investigaciones sobre conducta altruista, hay que decirlo, han dirigido su atención a descifrar, mediante más o menos complejos modelos matemáticos y simulaciones, el porcentaje de altruistas que debe haber en una población para que el altruismo subsista y viceversa, el porcentaje de egoístas que abocarían a los altruistas a la extinción y a la comunidad hacia la insolidaridad total. Este tipo de investigaciones, digamos más numéricas, acaso tengan su origen a partir de la aplicación por John

[5] Resúmenes sintéticos de dichas teorías se pueden encontrar en varios textos, aquí señalamos dos fácilmente accesibles a un público castellano: (Cela Conde y Ayala, 2005; Cela Conde et al., 2006).

Maynard Smith de la teoría de juegos al campo de la biología. Smith trataba de delimitar lo que llamó *estrategias evolutivamente estables* (EEE), es decir, aquellas estrategias que, desde el punto de vista adaptativo, son insuperables.

Ernst Fehr, a la postre científico social, toma un camino distinto al de otros estudiosos procedentes de ramas próximas a la biología (Gintis, Bowles, Boyd, Fehr, 2003; Fehr y Fischbacher, 2003; Fehr y Fischbacher, 2004). Fehr persevera en su apuesta por establecer comunicación entre la Economía y la Psicología, y centra sus indagaciones en el problema del altruismo humano, desde una perspectiva psicológica y sociocultural. En 2003 publica junto con su colega de la Universidad de Zurich, Urs Fischbacher, un artículo en que aborda el altruismo humano como un rasgo único en todo el mundo animal y en que, a la evolución genética, agrega la influencia de la evolución cultural. Hacen repaso de los conocimientos que actualmente tenemos del altruismo humano y que afirman, son mucho mayores que hace una década. Señalan que las sociedades humanas representan un caso anómalo en el mundo animal ejemplificado en rasgos como la compleja división del trabajo y la cooperación entre individuos no emparentados que, además, se dan en el seno de grupos muy numerosos.

Ambos autores afirman que el altruismo humano lleva mucho más lejos que el resto de animales, los mecanismos del altruismo recíproco y de la cooperación basada en la reputación (*reputation-based cooperation*)[6]. Tan lejos como hasta el concepto de reciprocidad fuerte (*strong reciprocity*) que sería una combinación de recompensa altruista (*altruistic rewarding*)[7], observancia de las normas (*norm-abiding behaviours*) y castigo altruista. La reciprocidad fuerte puede explicar la conducta cooperadora incluso en los casos en que la interacción no se repite y la posibilidad de ganar en reputación está ausente, ya que un individuo recíproco fuerte (*strong reciprocator*) recompensará a quien coopera y, por el contrario, castigará a quien no lo hace. Continúan Fehr y Fischbacher subrayando que para comprender correctamente la cooperación humana resulta esencial estudiar la interacción entre individuos egoístas e individuos recíprocos fuertes. Respecto al castigo altruista afirman que se trata de un mecanismo clave para hacer cumplir las normas sociales pero que el castigo no se debe tanto a la falta que recibe el propio castigador por parte del tramposo, cuanto a la falta que, digamos emblemáticamente, reciben los otros miembros del grupo en ese acto insolidario. Sin duda este matiz que aquí señalan los autores es el que venía justificando el apellido *altruista* en la

[6] La cooperación basada en la reputación puede entenderse como sinónimo, al menos parcial, de la reciprocidad indirecta.
[7] Recompensa altruista es la predisposición a premiar a otros por cooperar.

expresión castigo altruista. De lo contrario, no tendría sentido. Por último, Fehr y Fischbacher demuestran teóricamente que, dependiendo de las circunstancias, una minoría de altruistas puede forzar a una mayoría de egoístas a cooperar y viceversa, que una minoría de egoístas puede conducir a una mayoría de altruistas a comportarse de forma no cooperadora (Fehr y Fischbacher, 2003).

En 2004 Fehr y Fischbacher repiten colaboración en un segundo artículo en que plantean la necesidad de conocer los mecanismos subyacentes a las normas sociales (creación, cumplimiento, contenidos, etc.) para poder comprender el fenómeno de cooperación humana. Apuntan además que, si bien las sanciones son decisivas a la hora de hacer cumplir las normas, dichas sanciones pueden tener motivaciones no egoístas. La cooperación en las sociedades humanas, incluidas las actuales, se basa principalmente en normas sociales y en muchos casos dicha cooperación emana de normativas e instituciones legales. Desde este presupuesto los autores entienden imprescindible el estudio de las normas sociales para la mejor comprensión de la cooperación humana. En multitud de ocasiones las acciones de un individuo causan efectos positivos y negativos sobre los otros. Esto explica el interés de las personas en las acciones que realizan los demás, lo que frecuentemente se traduce en una demanda de normas sociales.

En este segundo artículo, los autores hablan de una norma social que denominan cooperación condicional (*conditional cooperation*) que estaría próxima a la estrategia *tit-for-tat* (estrategia del <<ojo por ojo>>) y acorde a la cual se comportaría la mayoría de la gente. En otras palabras, que un individuo aumentará su contribución a un bien público siempre y cuando las contribuciones de los demás miembros del grupo también se incrementen. Claro está, la contribución de unos y otros puede ser de mayor o menor cuantía, pero en los casos en que existe la posibilidad de sancionar a los bajos contribuyentes y a los tramposos, es presumible que los *cooperadores condicionales* cooperen voluntariamente en un alto nivel. El castigo, según los autores, es motivado más por un deseo de igualdad de resultados o de mera justicia que por el cálculo egoísta del interés personal. En conexión con lo anterior, Fehr y Fischbacher señalan que la capacidad humana de establecer y hacer cumplir las normas sociales tal vez sea la razón decisiva del carácter único de la cooperación humana (Fehr y Fischbacher, 2004). Esta capacidad humana para las normas sociales debe ser posible por el nivel de desarrollo cognitivo y emocional alcanzado únicamente por el ser humano en todo el mundo animal. Esta circunstancia habría permitido lograr con éxito altas cotas de cooperación. La selección cultural de grupo, por su parte, debe haber sido también muy relevante entre los humanos debido a sus capacidades cognitivas. El artículo concluye afirmando que las emociones juegan un papel importante en las

decisiones de cooperación y castigo, pero que todavía no se sabe si éstas -las emociones- dirimen tales decisiones o si, sencillamente, están relacionadas. Estos aspectos merecen ser estudiados para obtener mayor conocimiento de las condiciones en que surgen y logran estabilidad las normas.

En el número de agosto de 2004 de la revista *Science* aparece un artículo crucial en el desarrollo del concepto de castigo altruista. A mi entender, este artículo viene a reconocer la validez del concepto que será asumido desde entonces por un número creciente de investigadores. En el artículo titulado *The neural basis of altruistic punishment* (Quervain et al; 2004), de Quervain y colaboradores tratan de identificar, mediante el empleo de tecnología PET (tomografía de emisión positrónica), la base neural del castigo altruista. Parten de la hipótesis que el castigo altruista de tramposos produce alivio o satisfacción en el castigador y por tanto activará la región del cerebro relacionada con las recompensas. De los experimentos realizados, los autores sugieren que el caudado (*caudate*) juega un papel muy importante en el castigo altruista. El caudado, según los autores, es una región que ha sido relacionada con la toma de decisiones y la realización de acciones motivadas por la anticipación de una recompensa. De donde siguen que, la activación del caudado refleja la satisfacción anticipada de castigar a aquellos individuos que violan las normas.

Desde lo que podríamos llamar el ámbito de la Politología se le han opuesto alguna objeciones al concepto de castigo altruista (Fowler, Johnson, y Smirnov, 2004; Fowler, 2005; Egas y Riedl, 2005), pero ninguna de ellas parece atentar crucialmente contra el concepto. Por otra banda, en los últimos tiempos han aparecido algunas investigaciones que vienen a reforzar la idea de que ya sea por el castigo altruista, las instituciones sancionadoras o lo que algunos han dado en llamar castigo costoso (*costly punishment*), la cooperación en los grupos resulta mucho más estable en presencia de elementos represivos para los insolidarios o tramposos (Gürerk et al, 2006; Henrich, 2006; Henrich et al, 2006; Nakamaru e Iwasa, 2006).

3. UN CONCEPTO PROBLEMÁTICO.

3.1. ALGUNAS HIPÓTESIS.

El castigo altruista es un concepto difícil y todavía un tanto impreciso. No es un concepto evidente y desde luego aquel que no se halla familiarizado con los estudios sobre altruismo a duras penas puede advertir qué es aquello que se le refiere con la expresión *castigo altruista*. No obstante éste podría ser un rasgo asumible, pero

pronto se plantean otros inconvenientes. Cuando hablamos de castigo altruista, ¿de qué hablamos exactamente? ¿Se trata de un tipo de conducta altruista, un tipo de castigo, o acaso ambas cosas a la vez?

Castigo altruista es un concepto problemático y tal carácter le viene dado por dos circunstancias: a) se trata de un concepto complejo ya que es a un tiempo castigo y conducta altruista; b) se trata de un concepto "entre aguas" ya que se aproxima a categorías dispares como la venganza, la justicia, la ejemplaridad, el sacrificio o la benevolencia, sin identificarse plenamente con ninguna de ellas.

a) El castigo altruista es un tipo de castigo que presentaría las siguientes características:

-Más que de una acción, se trata de una reacción frente a los tramposos.

-El castigador es, con frecuencia, el propio individuo que recibe la falta.

-El castigador es consciente de su acción punitiva.

-El castigo es costoso para el castigador, aunque este coste siempre debería ser inferior a la pena impuesta al castigado.

-El castigo altruista puede cumplir, directa o indirectamente, varias funciones:

I. Corregir una conducta indebida.
II. Hacer cumplir la norma, ya sea ésta escrita o no.
III. Impedir que un individuo perjudique a los demás.
IV. Garantizar la cohesión del grupo, presentándose como un instrumento de control social.

-El castigo guarda cierta proporcionalidad con el grado de alejamiento (del tramposo) del nivel de cooperación esperado.

-El castigo es un acto, *a priori*, carente de publicidad.

b) Cuando afirmo que el castigo altruista es un concepto "entre aguas" me refiero a que con gran probabilidad se trata de un tipo de conducta con una importante base emocional, pero que toma elementos no de una sino de varias emociones tradicionalmente reconocidas. A la luz de los estudios realizados hasta la fecha no parece absurdo afirmar que el castigo altruista está próximo al sentimiento de venganza pero también al de justicia, ejemplaridad y benevolencia. O lo que es lo mismo, que el castigo altruista tiene algo de venganza pero no es venganza, tiene algo de ejemplaridad pero no es castigo ejemplar y podría tener algo de sentido de la justicia y de benevolencia pero no es plenamente ni lo uno ni lo otro.

El castigo altruista es un tipo de acción -o de reacción, como ya se ha dicho- causada por una emoción compleja o una concurrencia de emociones. Pero quisiera hacer notar que el castigo altruista es una acción altamente efectiva en tanto que expresión de los deseos y convicciones del propio actor. Ante un tramposo, un insolidario o un *gorrón*, el actor desea resarcirse de lo que considera una injusticia y, en cierta medida, vengarse de quien turba su sistema ético personal y/o quiebra la norma social. La satisfacción de ver cumplido este deseo superaría el coste de tener que sacrificar ciertos recursos en la ejecución de la acción.

Algunos aspectos del castigo altruista que podrían resultar de interés en futuras investigaciones serían los siguientes:

- Inspirado en una afirmación del filósofo inglés Jeremy Bentham (1748-1832), parece interesante indagar si aquellos individuos que se encuentran en una posición de mayor poder (dinero, autoridad, estatus, etc.) se muestran más desincentivados hacia la cooperación que otros.
- Parece fundamental acometer un estudio que permita cuantificar con cierta concreción el coste que asume el castigador en su acción punitiva y el coste que supone al tramposo el castigo. De tal manera que pudiéramos por una parte medir los límites del altruismo[8], si los hubiera, y por otra parte, corroborar la hipótesis de que el coste que soporta el castigador siempre es inferior al coste que soporta el castigado.
- Investigar en qué medida necesita el castigador comprobación ocular o garantía fiable de que el castigo que él ha decidido, se ejecuta. Este último punto está en relación con la hipótesis de que el castigo de los tramposos produce satisfacción al cooperador.

3.2. ¿HACIA ADÓNDE APUNTAN LOS TIROS?

Si el castigo altruista es aquella sanción que un individuo cooperador inflige, a costa de sus propios recursos (dinero, salud, bienestar, tiempo, etc.), a otro individuo no-cooperador y si se trata, pues, de una conducta con fuerte base emocional, cabe entonces preguntarse por el tipo de emoción de que se trata. En mi opinión el castigo altruista no consiste en ningún caso en una emoción desmedida, sino comedida, templada, y tal vez una emoción corregida por la cultura[9]. Esto es así al menos por dos circunstancias:

[8] Parte de esta empresa ha sido ya investigada por Fehr y Fischbacher (2003).

[9] <<la emoción corregida por la cultura>> es una expresión recogida de Carles Riba en el Prólogo al libro de Dantzer (1989).

la primera, que el castigador es consciente de su acción punitiva y libre para ejecutarla; la segunda, que el daño que sufre el castigado debería ser siempre mayor que el daño que se autoinflige el castigador.

La posibilidad abierta por el castigo altruista como mecanismo capaz de explicar la coevolución, biológica y cultural, de la cooperación humana ensancha el horizonte de preguntas que, sin duda, deberemos plantearnos en un futuro inmediato. Las conexiones entre biología y cultura son innegables pero sólo en la actualidad comenzamos a poder hablar científicamente de ello. Steven Pinker ha indicado cuatro nuevos campos de investigación que, en su opinión, podrán poner en relación la naturaleza y la sociedad. Dichos campos serían la Ciencia cognitiva, la Neurociencia[10], la Genética del comportamiento y la Psicología evolutiva[11]. Estas cuatro disciplinas <<aspiran nada menos que a aportar una explicación científica de la mente y de la naturaleza humana>>, y todo ello sin perjuicio de <<otras explicaciones más tradicionales en términos de aprendizaje, experiencia, cultura y socialización>> (Pinker, 2005).

[10] Estudio de las bases neurales del pensamiento, la percepción y la emoción.
[11] Estudio de la historia filogenético y las funciones adaptativas de la mente.

BIBLIOGRAFÍA.

En esta bibliografía se incluyen algunas referencias que no son citadas directamente en el trabajo pero de las que también nos hemos valido.

Ayala, F. J. (1999) “La teoría de la evolución”. Temas de hoy.

Ayala, F. J. y Cela Conde, C. J. (2006) “La piedra que se volvió palabra. Las claves evolutivas de la humanidad”. Alianza.

Barclay, P. (2006) Reputational benefits for altruistic punishment. *Evolution and Human Behavior*. **27**: 325-344.

Boyd, R. y Richerson, P. J. (1992) Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology*. **13**: 171-195.

Brosnan, S. F. y de Waal, F. B. M. (2003) Monkeys reject unequal pay. *Nature*. **425**: 297-299.

Campbell, D. T. (1983). The two distinct routes beyond kin selection to ultrasociality: implications for the humanities and social sciences. En Bridgeman, D. (1983) “The nature of prosocial development”. Academic Press.

Cela Conde, C. J. (1985) “De genes, dioses y tiranos. La determinación biológica de la moral”. Alianza.

Cela Conde, C. J. (¿?) Biology and altruism. How far is a human being from a bee?

Cela Conde, C. J. y Ayala, F. J. (2005) “Senderos de la evolución humana”. Alianza.

Cela Conde, C. J., Capó, M. A., Nadal, M. y Ramos, C. (2006) ¿Qué sabemos del cerebro social? Forum Barcelona. (En prensa)

Dantzer, R. (1989) “Las emociones”. Paidós.

Darwin, Ch. (2001) “El origen del hombre”. Edaf.

Dawkins, R. (2005) “El gen egoísta”. Salvat.

Egas, M. y Riedl, A. (2005) The economics of altruistic punishment and the demise of cooperation.

Fehr, E. y Gächter, S. (2000) Cooperation and punishment in public goods experiments. *The American Economic Review*. Vol. **90** No. **4**: 980-994.

Fehr, E. y Gächter, S. (2002) Altruistic punishment in humans. *Nature*. **415**: 137-140.

Fehr, E. y Fischbacher, U. (2003) The nature of human altruism. *Nature*. **425**: 785-791.

Fehr, E. y Fischbacher, U. (2004) Social norms and human cooperation. *TRENDS in Cognitive Sciences*. Vol. **8** No. **4**: 185-190.

Fowler, J. H. (2005) Altruistic punishment and the origin of cooperation. *PNAS*. Vol. **102** No.**19**: 7047-7049.

Fowler, J. H., Johnson, T. y Smirnov, O. (2004) Egalitarian motive and altruistic punishment. *Nature*. **433**: E1-E2.

Gandarias, J. M. y Hallet, D. (1989) "Basic english for the health sciences". Interamericana-McGraw-Hill.

Gintis, H., Bowles, S., Boyd, R. y Fehr, E. (2003) Explaining altruistic behavior in humans. *Evolution and Human Behavior*. **24**: 153-172.

Gürerk, Ö., Irlenbusch, B. y Rockenbach, B. (2006) The competitive advantage of sanctioning institutions. *Science*. **312**: 108-111.

Guzmán, R. A., Rodríguez-Sickert, C., Rowthorn, R. (2006) When in Rome, do as the Romans do: the coevolution of altruistic punishment, conformist learning, and cooperation. *Evolution and Human Behavior*. (En prensa).

Hamilton, W. D. (1964) The genetical evolution of social behaviour I & II. *Journal of Theoretical Biology.* **7**: 1-52.

Henrich, J. (2006) Cooperation, punishment, and the evolution of human institutions. *Science*. **312**: 60-61.

Henrich, J. et al. (2006) Costly punishment across human societies. *Science*. **312**: 1767-1770.

Henrich, J. y Boyd, R. (2001) Why people punish defectors? *Journal of Theoretical Bio*logy. **208**: 79-89.

Knutson, B. (2004) Sweet revenge? *Science*. **305**: 1246-1247.

Nagel, T. (1978) “The possibility of altruism”. Princeton University Press.

Nakamru, M. y Iwasa, Y. (2006) The coevolution of altruism and punishment: Rol of the selfish punisher. *Journal of Theoretical Biology*. **240**: 475-488.
12

Pinker, S. (2005) “La tabla rasa, el buen salvaje y el fantasma en la máquina”. Paidós.

Pratt Fairchild, H. (ed.) (2001). “Diccionario de Sociología”. FCE.

Price, M. E., Cosmides, L. y Tooby, J. (2002) Punitive sentiment as an anti-free rider psychological device. *Evolution and Human Behavior*. **23**: 203-231.

Quervain, D. J.-F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A. y Fehr, E. (2004) The neural basis of altruistic punishment. *Science*. **305**: 1254-1258.

Richerson, P. J. y Boyd, R. (1997) The evolution of human ultra-sociality. En Eibl-Eibisfeldt, I. y Salter, F. (eds.) (1997) “Ideology, Warfare and Indoctrinability”.

Smith, J. M. (1984) “La teoría de la evolución”. Hermann Blume.

Sober, E. y Wilson, D. S. (2000) “El comportamiento altruista. Evolución y psicología”. Siglo Veintiuno.

Trivers, R. L. (1971) The evolution of reciprocal altruism. *The Quarterly Review of Biology*. **46**: 35-57.

Turbón, D. (2006) “La evolución humana”. Ariel.

Wilson, E. O. (1980) “Sociobiología. La nueva síntesis”. Omega.

Yamagishi, T. (1998) “Trust and social intelligence: The evolutionary game of mind and society”. Tokyo University Press.